DOCUMENT RESUME

ED 163 082 TM 008 221

AUTHOR Willingham, Warren W.
TITLE Validity and the Graduate Record Examinations Program.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE 76
NOTE 36p.
AVAILABLE FROM Graduate Record Examinations, Educational Testing Service, Princeton, New Jersey 08541 (free while supplies last)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Admission Criteria; Bibliographies; College Admission; *College Entrance Examinations; *Graduate Study; Higher Education; Information Dissemination; *Institutional Research; Measurement Goals; *Research Needs; *Research Reviews (Publications); Testing Problems; *Testing Programs; *Test Validity
IDENTIFIERS *Graduate Record Examinations

ABSTRACT

The main purpose of this paper is to facilitate discussion of important issues concerning validity and to work toward a framework that the Graduate Record Examinations (GRE) Board Research Committee will find useful in assigning priorities and initiating projects. The background, scope, and meaning of the concept of the "validity" of the GRE are addressed in order to focus on the six proposed objectives for research on validity: (1) to encourage and facilitate institutional validity studies; (2) to deal effectively with methodological issues concerning validity that require the GRE program's initiative; (3) to develop improved criteria of success in graduate study; (4) to improve population validity and enhance understanding of it; (5) to improve institutional use of summary program data; and (6) to systematically insure the validity of revised or new measures resulting from program renewal. The six objectives are defined and the status of research relevant to each of them is presented in terms of reports available and current projects. (ROF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Educational Testing Service
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ED163082

TM008 221

GRE VALIDITY and the GRADUATE RECORD EXAMINATIONS PROGRAM

BY WARREN W. WILLINGHAM

EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY

Copyright © 1976 by Educational Testing Service. All rights reserved.

FOREWORD

In recognition of the importance of questions of validity to the Graduate Record Examinations program, the GRE Board asked Dr. Warren W. Willingham of the staff of Educational Testing Service to prepare a paper on the subject of validity and the GRE that could provide a basis for further Board discussion and decisions. After reviewing Dr. Willingham's paper, members of the Board agreed that it was an excellent document, which might well be of interest to others concerned with graduate admissions and the transition from undergraduate to graduate study. Accordingly, the GRE Board asked that the paper be published, and we are pleased to make it available.

Richard H. Armitage
Chairman, GRE Board

CONTENTS

INTRODUCTION 1

BACKGROUND OF THE PROBLEM 2
- Nature of the GRE program 2
- Previous research 3
- The current importance of research on validity 4

THE SCOPE AND MEANING OF "VALIDITY" OF THE GRE ... 7
- Conventional interpretations of validity 7
- Social interpretations of validity 8
- Construct validity of the program 10

SIX PROPOSED OBJECTIVES FOR RESEARCH ON VALIDITY .. 11
- Objective I: To encourage and facilitate institutional validity studies 11
- Objective II: To deal effectively with methodological issues concerning validity that require the GRE program's initiative 12
- Objective III: To develop improved criteria of success in graduate study 14
- Objective IV: Population validity: How to improve it and enhance understanding of it 17
- Objective V: To increase institutional use of summary program data 19
- Objective VI: To systematically insure the validity of revised or new measures resulting from program renewal 20

STATUS OF RESEARCH RELEVANT TO THE SIX OBJECTIVES 21
- Objective I: To encourage and facilitate institutional validity studies 21
- Objective II: To deal effectively with methodological issues concerning validity that require the GRE program's initiative 21
- Objective III: To develop improved criteria of success in graduate study 21
- Objective IV: Population validity: How to improve it and enhance understanding of it 22
- Objective V: To increase institutional use of summary program data 22
- Objective VI: To systematically insure the validity of revised or new measures resulting from program renewal 22

REFERENCES AND BIBLIOGRAPHY 23

INTRODUCTION

Validity is necessarily a major concern of any testing program. It is in the interest of the user that a test measure what it is supposed to measure and that it bear a reasonable relationship to the criteria it is intended to predict. It is the responsibility of a test sponsor to insure that these qualities prevail in the testing program. This principle of responsibility applies with a special force to a national program such as the Graduate Record Examinations (GRE), which affects large numbers of people. The stakes are high for both the individual and society.

Because of these considerations, the validity of examinations sponsored by the GRE program has always received close attention. It is suggested in this paper, however, that current issues concerning their validity are critical to the immediate future of graduate admissions. Therefore, it is also suggested that research on validity should have high priority in the GRE program over the next several years.

The main purpose of this paper is to facilitate discussion of important issues concerning validity and to work toward a framework that the GRE Board Research Committee will find useful in assigning priorities and initiating projects. Toward that end, subsequent sections of the paper provide background, define the scope of the problem, and outline six major objectives that might guide the Board's efforts in this important area of research.

BACKGROUND OF THE PROBLEM

To appreciate the various aspects of validity that apply to the GRE program, it is useful to consider the functions and role of the program, the kinds of research on validity undertaken to date, and why research on validity is especially important at this point in the life of the program.

Nature of the GRE program

The meaning one attaches to validity, as regards the GRE program, depends on how one perceives the role of the program. The GRE program is identified primarily as a series of examinations for use by graduate schools in selecting their students. The program has, however, a variety of functions and a variety of constituents. The following functions can be identified:

- to help with the admission of students to graduate school through publications, research, advisory services, and forum activities;
- to provide examination programs as measures of students' potential for success in graduate education;
- to provide an objective national basis for understanding the nature and distribution of academic talent by analyzing and describing characteristics of relevant student groups;
- to facilitate educational and career guidance by providing information to students and faculty;
- to inform undergraduate institutions what the graduate community considers adequate preparation for graduate study.

The following constituencies can be distinguished:

- students of different age, sex, background, and so forth;
- administrators and their institutions;
- different academic disciplines and fields;
- masters as well as doctoral programs.

There is considerable variation in the extent to which the GRE program serves these functions and constituencies, and there is, no doubt, considerable difference of opinion regarding their priority. Some have a recognized and explicit role in the program; the role of others may be largely implicit. For example, as the program has operated thus far, the roles of the last two functions listed above are more accurately characterized as potentially important than as of primary concern. Also, it seems likely that students comprise a more important constituency of the program in their own minds than in the minds of many institutional sponsors. Nonetheless, each of these functions and constituencies helps to define the responsibilities of the program.

In discussing the program's functions, it is necessary to distinguish between "the GRE program" and the "examination program." The "GRE program" denotes the entire program structure—its organization, governance, staffing, financing, research activities, major operational components, and so on. The "examination program" has a narrower meaning: the test or group of tests that a candidate may take or an institution require and the directly related services such as directions to examinees and information for interpreting their scores, guidelines for institutional use of the tests, analytic reports, and so forth. Unless stated otherwise, program is used hereafter to denote this narrower meaning.

Previous research

Research directly related to the validity of the examination program has fallen into one of two partially overlapping categories: (A) a variety of research carried out at ETS under GRE sponsorship and (B) validity studies carried out at other institutions. The overview which follows indicates the general types of GRE research that have a bearing on validity and the main conclusions concerning predictive validity that seem warranted on the basis of institutional studies. For reviews of methodological or other issues, see the reports cited below.

The relevant GRE research over the past 10 to 15 years has been partly concerned with traditional validity studies, but also it has treated a variety of other topics bearing upon validity. Part A of the attached bibliography lists 25 GRE publications that report pertinent research. They concern the following topics:

- methodological issues in the conduct of validity studies (Boldt, 1975; Reilly and Jackson, 1974; Rock, 1974, 1975)
- studies of the validity of the GRE in the selection of foreign graduate students (Harvey and Pitcher, 1963; Sharon, 1974)
- test bias and the use of the GRE in selecting minority applicants for graduate study (Echternacht, 1974; Flaugher, 1974)
- criterion problems and the analysis of what constitutes success in graduate study (Campbell, Freund, and Lannholm, 1965; Carlson, Evans, and Kuykendall, 1974; Reilly, 1974a, 1974b)
- studies concerning test use in selective admissions (Burns, 1970; Campbell, Hilton, and Pitcher, 1967; Burns, Dremuk, and others, 1971; Lannholm, 1962, 1968a; Madaus, 1966)
- special prediction studies and summaries of institutional validity studies (Lannholm, 1960, 1968b, 1972; Lannholm, Marco, and Schrader, 1968; Lannholm and Schrader, 1951; Olsen, 1955; Rock, 1972)

The institutional studies report on statistical analyses of the relationship between GRE tests and other predictors to various criteria of success in

graduate study. Willingham (1974) has provided the most recent analysis of all publications and reports of validity studies involving the GRE. Part B of the attached bibliography lists the 43 studies reported between 1952 and 1972 that Willingham analyzed in that article. These studies were based on 138 independent sets of data and 616 validity coefficients. The data indicated that:

- Validity coefficients for various predictors of graduate grade-point average (GPA) tend to be somewhat lower than corresponding coefficients at the undergraduate level. This is not surprising considering the restricted range of talent frequently encountered at the graduate level (see, for example, Dawes, 1975).
- The undergraduate GPA is a moderately good predictor of graduate GPA and faculty ratings; it is a poor predictor of whether a student will attain the Ph.D. Depending upon the success criterion used, the GRE composite of Verbal and Quantitative Ability scores is either slightly or substantially more valid than the undergraduate GPA.
- The GRE Advanced Test is the most generally valid predictor among those reviewed. It was typically more valid than the GRE Aptitude Test and had a higher validity than the undergraduate GPA in eight of the nine academic fields represented in the review.
- Recommendations are a fairly poor predictor of whether a student will successfully complete a doctoral program.
- A weighted composite including undergraduate GPA and one or more GRE scores typically provided a validity coefficient in the .40 to .45 range. This was somewhat higher than the validity of GRE scores alone and substantially higher than the validity of undergraduate GPA alone. This was the case for each success criterion and practically every academic discipline represented.

In addition to these empirical results, a variety of methodological and conceptual problems were cited that tend to create unusual difficulty in demonstrating the validity of entrance examinations at the graduate level. On the basis of these problems, the data available, and other considerations, Willingham concluded that (1) the efficiency of prediction is not likely to be enhanced merely through the development of improved predictors, and (2) the main hope for improved effectiveness in predicting success in graduate education lies in better definitions of what constitutes success, i.e., more reliable criteria that are more clearly differentiated with respect to training objectives.

The current importance of research on validity

Considering the number of graduate programs in the country and the far reaching importance of their admissions policies and procedures, few studies

have been made of the validity of the GRE for selecting graduate students. That fact and the fact that validity should always be a prime responsibility in any testing program are sufficient reasons for emphasizing research in this area. Furthermore, several current circumstances make validity a special concern of the GRE program. These circumstances follow three general themes.

First, the selection of graduate students is of greater concern now than in the past for the simple reason that many more students are involved. Selection often cannot be handled on a personal basis and, at the same time, the process is fragmented (typically along departmental lines) so that the statistical technology of selection frequently cannot be applied effectively. Concurrently, other trends are causing faculties to question the adequacy of selection practices. Undergraduate grades are assumed to be inflated and less trustworthy than in the past. New regulations to protect individual privacy give further reason to doubt the usefulness of personal recommendations. These developments suggest to some that GRE Verbal and Quantitative Ability scores should perhaps have greater weight in selection. At the same time there is increasing interest in the assessment of "competence" as opposed to aptitude. For example, it is now argued by some that selection in higher education should place more emphasis on traits that come closer to the real requirements of professional work (Hodgkinson, 1975). In support of this view, the modest relationship between college grades and adult success is frequently cited (Hoyt, 1965). All these developments and considerations contribute uncertainty as to what constitutes a valid basis for selecting students.

Second, these educational and methodological concerns are confounded by social and legal issues that have gained great importance in the last few years. To a considerable extent there is *de facto* acceptance of an egalitarian philosophy of admission in many institutions at the undergraduate level—at least in the public sector. In large part, admission to graduate study is still based upon merit, but this general rule is sharply conditioned by the widely perceived necessity to represent fairly those groups that constitute minorities in graduate education. This necessity raises complex questions concerning what constitutes unbiased selection when prediction is, as always, imperfect. The social issue becomes an important legal issue when the courts are asked to decide what constitutes a valid test and whether an institution must always select the student with the highest probability of success. Ironically, a decision either way is likely to raise questions of implementation that will require far greater sophistication concerning the validity of admission practices than presently exists. Whatever the resolution, when admission to privilege is treated as a legal issue, those responsible for the process must be able to defend its equity.

The third reason validity is currently such an important issue for the GRE program is that the Board is sponsoring a systematic research and development effort toward program renewal, i.e., shortening the Aptitude Test, developing additional modules for optional use, and examining ways

to make the program more useful to students and to institutions. Each of these efforts will require careful attention to the validity of proposed program changes. Not only must presently valid tests and procedures be maintained, the soundness of any new conceptions regarding valid measures and procedures in graduate admission must be demonstrated.

THE SCOPE AND MEANING OF "VALIDITY" OF THE GRE

Often the term validity is conceived narrowly—simply as a relationship between a test score and some measure of success in a subsequent activity. In considering what sorts of research on validity the GRE Board might want to undertake, it is necessary to take into account not only several conventional conceptions of validity, but also the fact that the program has various parts and various social implications.

Conventional interpretations of validity

The most common forms of validity are generally referred to as content validity, criterion-related validity, and construct validity. The definitions quoted in the following paragraphs are taken from *Standards for Educational and Psychological Tests* (American Psychological Association, 1974). The emphasis below has been added.

"Evidence of *content validity* is required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent. Content validity is most commonly evaluated for tests of skill or knowledge; it may also be appropriate to inquire into the content validity of personality inventories, behavior checklists, or measures of various aptitudes." Thus, content validity has special relevance to the Advanced Tests since these examinations must represent subject fields accurately and produce appraisals of knowledge that are fair regardless of the fact that undergraduate curriculums vary from institution to institution.

"*Criterion-related* validities apply when one wishes to infer from a test score an individual's most probable standing on some other variable called a criterion. Statements of predictive validity [for example] indicate the extent to which an individual's future level on the criterion can be predicted from a knowledge of prior test performance. . . . For many test uses, such as for selection decisions,. . . predictive validity provides the appropriate model for evaluating the use of a test or test battery." Predictive validity is central to the GRE program not only because the examinations are used to select students likely to succeed in graduate study, but also because there is increasing social and legal pressure against using tests for such purposes unless there is clear public evidence of such a relationship.

"Evidence of *construct validity* is not found in a single study; rather, judgments of construct validity are based upon an accumulation of research results. In obtaining the information needed to establish construct validity, the investigator begins by formulating hypotheses about the characteristics of those who have high scores on the test in contrast to those who have low scores. Taken together, such hypotheses form at least a tentative theory

about the nature of the construct the test is believed to be measuring." In considerable part, the construct validity of the GRE rests upon decades of psychometric research, indicating that verbal and quantitative ability play a critical role in most types of intellectual work, and upon even more extensive educational experience which indicates that frequently the best predictor of future success in an academic field is early competence indicated by a subject-matter test. Construct validation requires constant attention, however, to insure that a test is actually measuring the construct intended. For example, it is necessary to insure that a reading comprehension test is not so complicated in content as to stress reasoning instead of reading, or that a mathematics test does not use language that places a premium upon knowledge of vocabulary. Naturally, construct validation is even more demanding and important in the case of new measures in areas like cognitive style and creativity.

Social Interpretations of Validity

A number of broad interpretations of validity are associated with the social implications of test use. For example, there are such questions as the validity of a test for different groups of people, test validity as reflected in the ways test use affects the users, and longer range effects of using a particular test in a larger social context. These are more recent interpretations of validity. They deserve special consideration because the GRE program operates in an unusually broad social context.

First, a valid test must be fair and appropriate for all individuals taking the test. That is, it must be free of systematic bias and distortion vis-à-vis the various populations or subgroups taking the test, and the test should not have different meanings for such groups. Messick (1975) makes the important point that one validates not a test, but an interpretation of data derived from a specific procedure. Whether that procedure (test) has the same properties and patterns of relationships in different population groups is an important empirical question.

From a somewhat different angle, Thorndike and Hagen (1969) refer generally to validity as "whether the test measures what we want it to measure." Thus, while a test may be intended to measure knowledge of American history, there are a number of things it is *not* intended to measure; e.g., reading speed, cultural disadvantage, sex, age, language spoken in the home, and so on. In this sense there are an indefinite number of ways in which a test may be biased and an indefinite number of subgroups for which a test may not be appropriate. There is no way to guard against all such possibilities, and it can easily happen that making a test fairer for one person may make it less fair for another. It is evident, however, that bias and ap-

propriateness are important aspects of validity that require constant attention.

A second social interpretation of validity concerns the use and usefulness of a test in the context in which it is actually applied. "The casual phrase *test validation* seems to imply that the score one interprets comes from a naked instrument. The instrument however is only one element in a procedure and a validation study examines a procedure as a whole" (Cronbach, 1971). Tests are constructed with a purpose in mind and they are used in a context of instructions, interpretative materials, supporting research, expected effects, and constantly changing conditions of use. Again, validity does not reside in the test itself, but depends upon appropriate outcomes resulting from the use of the test. The GRE Board cannot be responsible for every conceivable instance of test misuse, but it should assume responsibility for making known the conditions of proper use, for advancing understanding of the social and educational implications of different uses of tests, and for insuring that tests are not presented in ways that might encourage inappropriate use.

There is another important relationship between validity and test use. It is especially true in a context like the GRE program that a test or measure does not stand alone. It competes for time and attention with other parts of the examination program that may be equally valid for the same purpose or more valid for other purposes. This leads to a third social interpretation of validity.

Cronbach (1971) distinguishes "educational importance" as a form of validity equal in stature, and parallel, to content validity and construct validity. He defines this form of validity as follows: "Does the test measure an important educational outcome? Does the battery of measures neglect to observe an important outcome?" The choice and content of measures included in the Board's examination program are significant because those measures constitute an important social communication. Regardless of whether they are intended as such, the GRE do to some extent communicate to colleges what the graduate community thinks colleges should teach and what students should learn. It is also argued that the GRE program needs to reflect the important learning outcomes of undergraduate education; i.e., it must follow the curriculum instead of leading it. From either point of view, the content of the program constitutes a message that has a bearing on education far beyond the admissions process. Consequently the inherent relevance, significance, and value of the traits measured deserve close attention.

These latter interpretations of validity bear especially upon the usefulness and appropriateness of different components of the examination program, both as they serve the immediate purposes for which they are intended and as they may be justified in some broader educational sense. Obviously such interpretations do not apply to isolated measures, but to the program as a whole. These considerations and the foregoing discussion suggest a broader notion—one of program validity as distinct from test validity.

Construct validity of the program

Principles concerning validity can be applied to tests and other individual measures, to batteries or groups of measures, or to an integrated program that may have a variety of pieces (e.g., a central core, a variety of test options, biographical information, special measures, and so on). Each of those pieces should have a rationale concerning its legitimate and useful function. But the pieces are not free-standing. The various components are used in a context of related procedures, materials, services, and so on. Each part of the program is to some extent dependent upon other parts and upon an overall rationale as to how the program serves its functions and its constituents.

These considerations suggest that the Board should be guided by an overarching sense of the *construct validity of the program*. In this context, a valid test is a *defensible* test; i.e., an accurate and fair measure of what you want to measure and also one that is useful for its purpose. More specifically, the notion of the construct validity of a program suggests that a test or measure is a valid component of a program if it meets these conditions:

1. It represents fairly what is intended. That is, it satisfies concerns such as content validity, construct validity, educational importance, and appropriateness for the examinees, both as one group and as subgroups.
2. Its use is demonstrably effective. It meets the requirements of criterion related validity, predictive bias, characteristics of the program that affect test use, and legal issues concerning test use.
3. It serves a distinctive purpose in relation to other tests and measures in the program, that is a purpose not served by the other tests and measures.

SIX PROPOSED OBJECTIVES FOR RESEARCH ON VALIDITY

The previous discussion provides background for the problem. In this section six research objectives are suggested in order to provide for the GRE Board Research Committee's consideration specific proposals for action. The objectives are as follows:

I. To encourage and facilitate institutional validity studies
II. To deal effectively with methodological issues concerning validity that require the GRE program's initiative
III. To develop improved criteria of success in graduate study
IV. Population validity: How to improve it and enhance understanding of it
V. To improve institutional use of summary program data
VI. To systematically insure the validity of revised or new measures resulting from program renewal

In the following paragraphs an initial statement of each objective is followed by a brief rationale and discussion of several issues relevant to the objective. These issues are discussed either as general research needs or, in some cases, as more specific possible projects. But the main purpose is to suggest a framework for thinking about validity research that is needed.

Objective I: To encourage and facilitate institutional validity studies

The American Psychological Association (1974) outlines a variety of responsibilities of test sponsors for examining and establishing the validity of measures they offer for use. In this paper we give special attention to these responsibilities of the GRE program, but the conditions under which tests may be valid or invalid are essentially unlimited because applications vary so widely with respect to purpose, academic field, criteria, local conditions, and so on. The GRE Board cannot hope to establish validity in even a significant minority of the possible situations in which the tests may be used.

Consequently, it is important for users to recognize their own responsibility for examining the validity of a test for the purpose and circumstances they have in mind. As Cronbach (1971) states, "In the end, the responsibility for valid use of a test rests on the person who interprets it. The published research merely provides the interpreter with some facts and concepts. He has to combine these with his other knowledge about the persons he tests and the assignments or adjustment problems that confront them, to decide what interpretations are warranted." But users confront many problems in carrying out institutional validity studies. In most cases the appropriate locale is the individual department where, however, the number of students may be small and the faculty may lack sufficient interest or expertise to

pursue the question of validity. It is important, therefore, that the GRE program find ways to encourage and facilitate institutional validity studies.

One possibility would be to develop a limited program of cooperative validity studies. This might involve identification of individual departments or institutions where there is the interest and possibility of carrying out a study of more than routine interest. With technical advice from staff at Educational Testing Service (ETS), the institution might organize and supply appropriate data; a researcher in Princeton might analyze the data and prepare a report, probably in some model format. This process might also involve identification and announcement of priority areas of interest; e.g., particular academic fields, interesting possibilities for criterion development, special populations of students, and institutionwide use of program information. Periodically such group studies might be collected and reported.

Another possible approach would be to develop a validity study kit that might consist of a step-by-step notebook for doing local studies, useful references and forms, a collection of relevant reprints, and so on. This model for encouraging local studies has the virtue of cost-effectiveness, but it also places most of the responsibilities for initiative on the institution. ETS's main responsibility would be to produce the kit. That itself will require experience and good ideas if such a kit is to be useful and cope effectively with the variation of local circumstances.

Another research need somewhat related to those above is the desirability of developing effective relationships with institutions in order to facilitate work in this area. On the one hand, both of the above possibilities can be greatly facilitated by working intensively with one or two institutions over a reasonable period of time to explore the problems of conducting institutional validity studies at the graduate level. Furthermore, there will likely be a need to develop a cooperative relationship with a variety of institutions in order to validate experimental modules that may be considered for inclusion in the GRE program over the next several years. This need is directly related to Objective VI, though the development of the necessary institutional relationships will profit from groundwork prior to the time when the need actually arises.

Objective II: To deal effectively with methodological issues concerning validity that require the GRE program's initiative

A variety of traditional issues typically referred to as technical problems make validity studies exceptionally difficult at the postgraduate level. Perhaps the most serious is the criterion problem, though it is a transcendent issue concerning conception as well as methodology and deserves separate

discussion as Objective III below. The most familiar technical issues concern very small samples, often severe restriction in the range of talent due to selection, and lack of confidence in the meaning and reliability of undergraduate grades. The following paragraphs outline some of these issues and suggest some possibly fruitful lines of research.

Validity studies carried out in individual departments are often based upon small samples and a very restricted range of test scores and undergraduate grades. These conditions often combine to produce low and erratic validity coefficients. A related but different view of this problem is the fact that there has been very little attention given to the validity of the GRE among departments within individual disciplines or fields. It would appear desirable to give additional attention to ways of pooling data across departments, thereby mitigating the technical problems and also demonstrating validity in a larger context. This requires use of some common criterion. Perhaps the only one that would make sense is some general notion of success in graduate education, such as completion of the degree or overall faculty ratings.

Another general possibility for dealing with these issues is a retrospective nomination study; i.e., asking faculty in a number of departments within a field to nominate outstanding and poor students over a period of several years. It may be possible to develop substantial samples for studies within selected fields. This type of study would also require a common criterion of success, such as obtaining the Ph.D.

A special advantage of these two types of studies is the fact that they can be carried out over a limited time period. But, in addition to narrowly conceived validity studies, many especially interesting research questions require longitudinal study. Serious consideration should be given to the development of a longitudinal study that follows a carefully structured sample of students through and beyond graduate education. Students from several fields might be included with oversampling of special groups of interest. With several spaced follow-ups, a group of cooperating students could provide a valuable data base for a variety of studies in addition to specific investigations designed at the outset. Topics of special interest in such a longitudinal study would include follow-up of minority students through and beyond graduate education, studies of patterns of attendance and career choice, and analysis of the cost of graduate education and how financing alternatives affect students' decisions.

The quality of undergraduate grades as a predictor is another familiar problem. It is commonly understood and accepted that a "B" at one institution is not necessarily equivalent to a "B" at another institution. A good deal of research at the undergraduate level has indicated that there is no value in trying to adjust grades from different high schools if admissions decisions are made on the basis of grades and an entrance examination. At the graduate level, however, it seems that corrections for variations in grading standards from one undergraduate college to another can sometimes improve predictions slightly. Pitcher and Schrader (1972) report that multiple

correlations predicting success in graduate business schools are increased on the average about .02 when the quality of the undergraduate institution is taken into account. A similar result has been found in the case of law school admissions. This matter may be worth an exploratory investigation in connection with the GRE, though it is not clear whether the potential gain would make such an investigation worthwhile, nor whether it would be easy to implement the results in any event. Furthermore, applying institutional "corrections" may involve difficult political problems in graduate admissions.

A more serious question may be the status of undergraduate grades in general. Informal opinions suggest a mounting lack of confidence in the reliability and dependability of undergraduate grades as a predictor due to the fact that grades are severely inflated and the fact that some faculty are opposed to competitive grading as a matter of principle. It is difficult to judge the extent to which undergraduate grades have actually been seriously compromised as a measure of accomplishment and predictor of subsequent academic success. A careful study of college grading, perhaps with resulting recommendations from the Board, could possibly be of great value in reversing an undesirable trend or, at the very least, in revealing the character and scope of the problem and suggesting ways of dealing with it.

A related issue is the problem of interpreting the credentials of undergraduate students when they do not come in the form of grades at all. An increasing number of institutions are recognizing various forms of nontraditional learning, converting to competency-based curriculums, awarding credit for experiential learning, and experimenting with narrative transcripts. A study of how such credentials are evaluated and to what extent they forecast success in graduate study could provide a valuable service by helping to influence evolving nontraditional practices in sound directions.

Objective III: To develop improved criteria of success in graduate study

It is commonly acknowledged that a principal difficulty in establishing the validity of predictors and procedures for admitting graduate students is the lack of clear-cut reliable criteria of success. Grades are widely suspect; methods of evaluating students' performance in graduate education vary greatly from department to department; and the ultimate criterion of degree attainment takes many years to ascertain and depends upon many intangible and fortuitous events. To a considerable extent the criterion problem is due to the very common tendency of those responsible for educational selection programs to focus upon predictors and to ignore the measures of achievement that those predictors are intended to forecast.

The need to give specific attention to the rationale and measurement (i.e., construct validity) of criteria is well-stated by Cronbach (1971): "The

asymmetric conception of the test as a predictor of a certain performance has been discarded in favor of a symmetric view. According to this, persons are observed in situations. Some are artificial occasions for observations, which are called tests, and some are situations arising in the natural course of the person's work or schooling. Relating these observations to each other tells one about the situational demands and about the resources individuals bring to bear. To study the validity of a test interpretation is to study how behavior in one situation is related to behavior in another. Both observations reveal characteristics of the individual, and both types of behavior should be understood."

Defensible and reliable criteria of success are likely to become more important in the face of lagging confidence in grades and the increasing need to justify administrative actions, both with respect to admitting and dropping students. Furthermore, admission standards are likely to come under increasing scrutiny, partly by those speaking for underrepresented groups who question the social equity of current practices and partly by those who assert there is undue reliance upon aptitude tests and objective measures in general. Should legal action require empirical justification of admission decisions, the need to develop sound criteria will immediately become critical.

Willingham (1974) has urged much greater attention to the problem of criteria, especially in the context of a broader view of predictor-criterion relationships and alternate strategies of selection. The alternate strategies depicted in Figure 1 imply that different departments or programs within departments may emphasize different training objectives, which in turn should be related to the way students are selected and the way their performance is evaluated.

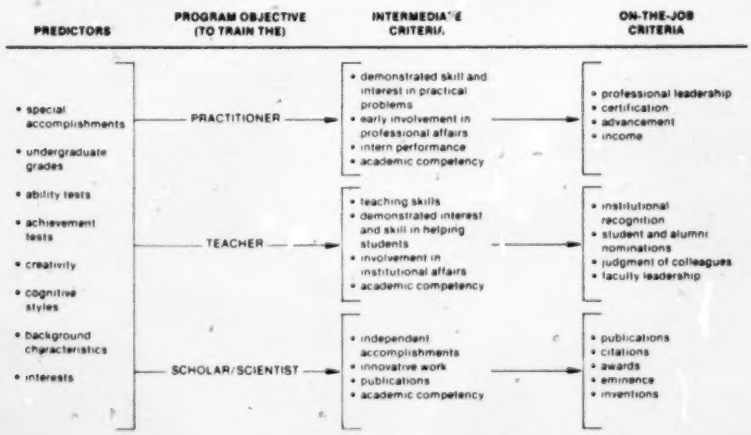

Figure 1: Alternate prediction strategies in graduate education. Reproduced with permission from *Science*, 1974, 183, (4122), 277. Copyright 1974 by the American Association for the Advancement of Science.

Another aspect of Figure 1 deserves special attention. Moving from left to right in the figure, *predictors, intermediate criteria,* and *on-the-job criteria* form a prediction chain. Preadmission predictors such as test scores or special achievements are intended to predict success in graduate school (an intermediate criterion). Performance in graduate school should, in turn, predict success in subsequent career-related activity. Since ultimate career success depends upon many circumstances and developed skills not necessarily related to performance in graduate school or to earlier predictors, it is improper to think of a graduate admission test as a predictor of job success. On the other hand, if a test or other measure is used to screen students, it ought to be helpful in predicting success in school and also should have a logical relationship to long-range criteria. That is, the test should have *construct validity* in the sense that it represents an ability of demonstrable importance (perhaps one of many) in determining long-range success. Developing a better conception of intermediate criteria of success in graduate school is an important step in examining the logical relationships in the prediction chain.

There are several strategies that might be helpful in developing improved criteria. It seems especially important to encourage systematically the development of better criteria in the routine execution of validity studies as well as the routine evaluation of students' performance. This might be accomplished partly by working with individual institutions to carry out and publish model studies that can help to illustrate the development of different types of criteria. For example, it would be desirable to illustrate and encourage the use of reliable criteria like rating scales, such as those developed through GRE research (Carlson, Reilly, Mahoney, and Casserly, 1975), or comprehensive examinations, which may already be available within departments but are not normally used in validity studies.

Considering the importance of criteria and the critical role they could play given additional legal interest in graduate admissions, it may be desirable to undertake a fairly systematic analysis of how graduate departments view success. A useful follow-up to the Carlson, Evans, and Kuykendall (1974) survey would be an intensive analysis of the rationale and basis upon which departments evaluate their students, the evaluation procedures actually employed, and the psychometric properties of the resulting criteria. This sort of analysis could have considerable value in describing how high level talent is currently assessed and in suggesting opportunities and possible road blocks in the development of improved criteria.

Another generally desirable strategy would be to foster the development of such intermediate criteria as depicted in Figure 1. This might involve faculty ratings of particular types of accomplishments, special means of collecting outside judgments, or whatever procedures may be required to obtain information that is relevant to the most important training objectives. One way of encouraging the development of such criteria is through the cooperative institutional studies described under Objective 1. Another approach is a special developmental project which may be required in the

case of an unusually complex criterion such as scientific creativity. The present GRE research on scientific creativity is concerned specifically with the development of an intermediate criterion.

Another potentially useful approach starts with the observation that practically all validity studies incorporate criteria based upon student performance in graduate school. This typical design has two shortcomings if one is interested in confirming the "ultimate" social relevance (i.e., construct validity) of the GRE. First, typical validity studies are not directly relevant to the question of whether the GRE are effective in selecting for graduate study people who are likely to reach the highest levels of professional success. While screening prospective professionals from a group of graduate students is primarily the responsibility of the graduate schools, tests used for earlier screening should not be counterproductive in that process; that is, one would like to know that very successful professionals have typically scored well so that screening out students with low scores is both efficient and defensible. Second, studies that use success criteria relative to the standards of individual institutions may seriously underestimate the usefulness of the examinations because the range of talent is typically restricted at individual institutions, but ranges widely from one institution to another. Thus, the GRE may be relatively poor predictors of graduate performance in a single prestigious history department, but may provide a reasonably good indication of differential competence among graduates of all history departments.

It might be worthwhile, therefore, to examine the feasibility of determining Aptitude Test score levels for pertinent groups of individuals who have achieved some formal measure of success in their field. These might include such *ad hoc* groups as fellows of learned societies, officers of professional organizations, faculty of prestigious departments, individuals listed in honorific biographies, and so on. Comparison of the scores of such individuals with appropriate normative groups would be interesting, even admitting the possibility that some individuals may achieve prestigious status partly because at one time they were known to have scored highly on the test, or in spite of having scored poorly.

Objective IV: Population validity: How to improve it and enhance understanding of it

Messick and Barrows (1972) used the term *population validity* in referring to the generalizability of research findings across different populations. A similar notion applies to the validity of tests and other psychometric measures. If a test leads to incorrect inferences about a particular population, then the test is to that extent invalid. Incorrect inferences may result from the fact that the test itself is not a good measure for that population or that the test does not have the same relationship with the criterion for that popu-

lation. We can refer to these two sources of error as appropriateness and predictive bias.

Research on test appropriateness has been limited largely to item-group interaction studies and has not been especially productive as yet. More attention has been directed to predictive bias and most such studies have compared predicted and actual performance of black and white students. Linn's (1973) review of that research suggests a small but fairly consistent tendency for tests to overpredict the college grades of black students. Recently, attention has been directed to technical factors such as reliability that might explain that finding (Linn, Note 1).

A significant recent development has been the systematic analysis of the psychometric characteristics and social implications of different selection models (Petersen and Novick, 1974). Somewhat to the surprise of many people working in this field, it is now apparent that there are different selection approaches with rather different implications that are complex both technically and politically. For example, it can be demonstrated that a selection model that is fair to an individual may not be fair to a social group to which that individual belongs.

These different selection models may give rise to legal complications and, if so, perhaps even greater complications for the GRE program and for admission committees. The resolution of these issues may be largely political, but it seems important for the Board to pursue relevant research and other activities in this area that will help to illuminate the issues and clarify the implications for the program. Several possibilities can be suggested.

The program has undertaken several studies to analyze item-group interactions in an effort to identify types of items that might be unfair to some populations of examinees. That work has not been fruitful in discovering significant numbers of such items or in developing effective ways of identifying an unfair item prior to its use. More recent analysis of these types of data reveal some small but interesting tendencies for certain types of items to be harder or easier for different subgroups of examinees. In most cases it seems doubtful that these differences can be termed bias in any reasonable sense of the word, but the line of inquiry is useful from a research standpoint and desirable from the standpoint of monitoring the fairness of the GRE.

A promising and closely related line of research is the possibility of a more general attack on the question of the appropriateness of a test for the various populations of people who might take it. The Board has underway a developmental project that may succeed in creating "appropriateness indices" that would permit study of this problem from a different vantage point. The work is largely mathematical and, if successful, will likely take several years to improve our understanding of population validity.

While the Board must be concerned about the possibility of predictive bias in relation to any population of examinees, a special problem exists with respect to individuals who have been out of the educational system for an extended period of time. There is a common assumption that an objective examination is to some extent unfair to an individual of 35 who has been

away from college for a dozen years. In considerable part this problem parallels the issue of test bias with respect to minorities and can be studied in a similar fashion.

A substantial literature has developed concerning the technical, educational, and legal issues surrounding the general matter of population validity. There are reviews of various aspects of test bias, legal decisions, and implications for admissions committees. These problems are of considerable concern to departmental faculties, though the literature is likely not well known or available to them. The Board might consider whether the publication of a limited group of reprints with commentary would be a useful service beyond its present activities in this area.

Objective V: To improve institutional use of summary program data

Validity is often conceived of only in relation to the individuals who take a test. But test performance data are often reported about groups of examinees and inferences are drawn concerning those groups and the educational programs in which they have taken part. As the GRE program seeks to serve better the interests and needs of institutional sponsors, such summary data should be reported more frequently and systematically. The prospect of more systematic reporting of program data intensifies the need to insure that such data serve the useful purposes intended and do not foster erroneous inferences. Two examples illustrate the opportunity and the need.

The GRE program generates biographical and test data that can help institutions to understand the dynamics of the admissions process and possibly alter the process to serve their institutional purposes. The flow of talent into and through graduate education can be conceived of as a diminishing pool of individuals; viz., the pool of applicants, the group offered admission, the group who actually enroll, and finally the group who obtain a degree. Data from the program can be used not only to describe the characteristics of these successively diminishing groups, but also to identify the personal and demographic characteristics that operate most strongly in the selection process preceding each stage. Such analysis should help educational planners make useful connections between resource allocations, program planning, and encouragement of talent.

As a second illustration: the Board has taken steps to relate the GRE program to the Undergraduate Assessment Program at ETS. One aspect of this new service is to develop additional subscores for the GRE Advanced Tests and report them in summary form for groups of students in undergraduate departments. This sort of reporting can be quite useful to departments if it emphasizes the comparison of departmental objectives with student performance on corresponding parts of the examinations. The design of an interpretive framework of this sort could be an important contribution of

the GRE Board in improving cooperation between undergraduate and graduate education.

Objective VI: To systematically insure the validity of revised or new measures resulting from program renewal

In recent meetings the GRE committees and the GRE Board have given careful consideration to a broad effort aimed at program renewal over the next several years. Naturally this effort is experimental and highly tentative, but it does include a number of possible changes or additions to the GRE program—among them a possible shortening of the Aptitude Test and the development of optional modules that might be incorporated in the test, improvements in the Advanced Tests, instruments that might be useful for purposes other than selective admissions, and improved ancillary services for institutions and students.

It is impossible to know which of these promising possibilities will eventuate, but the process of program renewal places a special responsibility on the Board to insure the continuing integrity and validity of the program. There is firm consensus that renewal is important and that the validity of the program will thereby be enhanced, but a complex national program has rippling effects and relationships that are often difficult to anticipate. Consequently, change should not be taken lightly. It should be taken only on the basis of reasonable assurance that revised or new components have demonstrable validity as they were intended. Such assurance should be a conscious objective of the Board. Specific necessities for validation will arise as new program components are developed. Possibilities already apparent include the following:

- examination of the factorial and predictive validity of new versus old forms of the verbal and quantitative Aptitude Test;
- development of a normative framework for the interpretation and possible operational use of a measure of cognitive style;
- construct and criterion-related validation of objectively scored intermediate criteria of scientific creativity;
- content and predictive validation of an inventory of documented accomplishments.

These six objectives do not exhaust the range of important priorities concerning validity, nor are the possibilities mentioned under each objective intended to be inclusive. They are, however, indicative of the very diverse issues that require attention if the GRE Board is to feel reasonably confident that its examination program does exhibit construct validity in all essential respects. In that spirit the foregoing should be viewed as a possibly useful framework and orientation, not a prescription.

STATUS OF RESEARCH RELEVANT TO THE SIX OBJECTIVES

The following outline lists under each objective the relevant current GRE research as well as projects conducted and reported upon in the past As suggested by the previous discussion, however, there is need for further research of various types. Some has already been proposed to the GRE Board Research Committee, some is still in the realm of possibility.

Objective I: To encourage and facilitate institutional validity studies

Reports

- Predicting graduate school success (Lannholm and Schrader, 1951)
- Predicting success in Yale School of Forestry (Olsen, 1955)
- Abstracts of selected validity studies (Lannholm, 1960)
- Review of validity studies (Lannholm, 1968a)
- Cooperative validity studies (Lannholm, Marco, Schrader, 1968)
- Summary of validity studies (Lannholm, 1972)
- Predicting success in graduate education (Willingham, 1974)

Current Project

- Cooperative institutional validity studies (Wilson)

Objective II: To deal effectively with methodological issues concerning validity that require the GRE program's initiative

Reports

- The test chooser (Rock, 1974)
- Effects of option weighting on validity (Reilly and Jackson, 1974)
- Population moderators (Rock, 1975)
- Bayesian and least squares prediction (Boldt, 1975)
- Prediction of Ph.D. attainment in psychology, mathematics, and chemistry (Rock, 1972)

Objective III: To develop improved criteria of success in graduate study

Reports

- A study of the Advanced History Test (Campbell, Freund, and Lannholm, 1965)
- Critical incidents of graduate performance (Reilly, 1974a)
- Factors in graduate performance (Reilly, 1974b)

- Feasibility of common criterion validity studies (Carlson, Evans, and Kuykendall, 1974)
- Criterion rating scales (Carlson, Reilly, Mahoney, and Casserly, 1976)

Current Projects
- Intermediate criteria of creativity (Frederiksen and Ward)
- How graduate departments assess students (Wilson)

Objective IV: Population validity: How to improve it and enhance understanding of it

Reports
- Prediction of graduate performance of foreign students (Harvey and Pitcher, 1963)
- Use of TOEFL and GRE in predicting success of foreign students (Sharon, 1974)
- A quick method of determining test bias (Echternacht, 1974)
- New definitions of test bias (Flaugher, 1974)

Current Projects
- Development of an appropriateness index (Levine)
- Guessing instructions and subgroup performance (Pike)

Objective V: To improve institutional use of summary program data

Current Projects
- Development of additional psychology subscores (Altman and McPeek)

Objective VI: To systematically insure the validity of revised or new measures resulting from program renewal

Current Projects
- Cognitive style longitudinal study (Witken and Ward)
- Machine-scored field dependence test (Reilly and Donlon)
- Validity of scientific thinking tests (Frederiksen)
- Inventory of accomplishments (Baird)
- Machine-scorable tests of scientific thinking (Frederiksen and Ward)
- Factor analytic study of current and proposed forms of the GRE Aptitude Test (Powers, Swinton, and Carlson)
- Research into shortening V and Q and adding a new measure to the Aptitude Test (Conrad and Altman)

REFERENCES AND BIBLIOGRAPHY

Note that GRE Board publications with a P in the publication number are of interest primarily to researchers.

References

American Psychological Association. *Standards for educational and psychological tests.* Washington, D.C.: APA, 1974.

Carlson, Alfred B., Evans, Franklin R., and Kuykendall, Nancy M. *The feasibility of common criterion validity studies of the GRE* (GRE Board Professional Report 71-1P). Princeton, N.J.: Educational Testing Service, July 1974 (ERIC Document Reproduction Service No. ED 097 367). Previously published as Research Memorandum RM-73-16, July 1973.

Carlson, Alfred B., Reilly, Richard R., Mahoney, Margaret H., and Casserly, Patricia L. *The development and pilot testing of criterion rating scales.* (GRE Board Professional Report 73-1P). Princeton, N.J.: Educational Testing Service, October 1976.

Cronbach, Lee J. Test validation, in Thorndike, Robert L. (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Dawes, Robyn M. Graduate admission variables and future success. *Science,* 1975, 187 (4178): 721-723.

Hodgkinson, Harold. Cited in NIE's new director questions credentialing. *Education Daily,* June 24, 1975, p. 3.

Hoyt, Donald P. *The relationship between college grades and adult achievement: A review of the literature* (ACT Research Report Number 7). Iowa City, Iowa: American College Testing Program, 1965. (ERIC Document Reproduction Service No. ED 023 343)

Linn, Robert L. Fair test use in selection. *Review of Educational Research,* 1973, 43 (2): 139-161.

Messick, Samuel. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist,* 1975, 30 (10): 955-966.

Messick, Samuel, and Barrows, Thomas S. Strategies for research and evaluation in early childhood education, in *National Society for the Study of Education, Seventy-first yearbook, Part II.* 1972, pp. 261-290.

Petersen, N.S., and Novick, M.R. *An evaluation of some models for test bias* (ACT Technical Bulletin No. 23 and Iowa Testing Program Occasional Paper Number 8). Iowa City, Iowa: American College Testing Program, September 1974.

Pitcher, Barbara, and Schrader, William B. *Indicators of college quality as predictors of success in graduate schools of business* (Admission Test for Graduate Study in Business Brief Number 6). Princeton, N.J.: Educational Testing Service, 1972.

Thorndike, Robert L., and Hagen, Elizabeth. *Measurement and evaluation in psychology and education* (3rd ed.). New York: John Wiley and Sons, 1969.

Willingham, Warren W. Predicting success in graduate education. *Science*, 1974, 183 (4122): 273-278.

Reference note

Note 1.
Linn, Robert L. *Test bias and the prediction of grades in law school.* Paper prepared for the Conference on the Future of Law School Admission Council Research, September 27-28, 1974.

Bibliography—Part A

Boldt, Robert F. *Comparison of a Bayesian and a least squares method of educational prediction* (GRE Board Professional Report 70-3P). Princeton, N.J.: Educational Testing Service, June 1975. Previously published as Research Bulletin RB-75-15, April 1975.

Burns, Richard L. *Graduate admissions and fellowship selection policies and procedures, Part I* and *Part II* (a report of the GRE Board). Princeton, N.J.: Educational Testing Service, 1970.

Burns, Richard L., Dremuk, Richard, Hein, Andrew J., Robey, Richard C., Scudder, Hervey C., and Wiltsey, Robert G. *Graduate admissions and fellowship selection policies and procedures: Case studies* (a report of the GRE Board). Princeton, N.J.: Educational Testing Service, 1971. Out of print. (ETS Archives Microfiche #491)

Campbell, Joel T., Freund, Lucy, and Lannholm, Gerald V. *A study of performance on the Advanced History Test* (GRE Special Report 65-2). Princeton, N.J.: Educational Testing Service, August 1965.

Campbell, Joel T., Hilton, Thomas L., and Pitcher, Barbara. *Effects of repeating on test scores of the Graduate Record Examinations* (GRE Special Report 67-1). Princeton, N.J.: Educational Testing Service, April 1967.

Carlson, Alfred B., Evans, Franklin R., and Kuykendall, Nancy M. *The feasibility of common criterion validity studies of the GRE* (GRE Board Professional Report 71-1P). Princeton, N.J.: Educational Testing Service, July 1974 (ERIC Document Reproduction Service No. ED 097 367). Previously published as Research Memorandum No. 73-16, July 1973.

Echternacht, Gary. *A quick method for determining test bias* (GRE Board Professional Report 70-8P). Princeton, N.J.: Educational Testing Service, July 1974 (ERIC Document Reproduction Service No. ED 067 404). Previously published as Research Bulletin RB-72-17 April 1972.

Flaugher, Ronald L. *The new definitions of test fairness in selection: Developments and implications* (GRE Board Research Report 72-4R). Princeton, N.J.: Educational Testing Service, May 1974 (ERIC Document Reproduction Service No. ED 097 336). Previously published as Research Memorandum RM-73-17, September 1973.

Harvey, Philip R., and Pitcher, Barbara. *The relationship of Graduate Record Examinations Aptitude Test scores and graduate school performance of foreign students at four American graduate schools* (GRE Special Report 63-1). Princeton, N.J.: Educational Testing Service, April 1963.

Lannholm, Gerald V. *Abstracts of selected studies on the relationship between scores on the Graduate Record Examinations and graduate school performance* (GRE Special Report 60-3). Princeton, N.J.: Educational Testing Service, November 1960. Out of print. (ETS Archives Microfiche #236)

Lannholm, Gerald V. *The use of Graduate Record Examinations in appraising graduate study candidates.* (GRE Special Report 62-3). Princeton, N.J.: Educational Testing Service, October 1962.

Lannholm, Gerald V. *Review of studies employing GRE scores in predicting success in graduate study, 1952-1967* (GRE Special Report 68-1). Princeton, N.J.: Educational Testing Service, March 1968a.

Lannholm, Gerald V. *The use of GRE scores and other factors in graduate school admissions* (GRE Special Report 68-4). Princeton, N.J.: Educational Testing Service, October 1968b.

Lannholm, Gerald V. *Summaries of GRE validity studies 1966-1970* (GRE Special Report 72-1). Princeton, N.J.: Educational Testing Service, February 1972.

Lannholm, Gerald V., Marco, Gary L., and Schrader, William B. *Cooperative studies of predicting graduate school success* (GRE Special Report 68-3). Princeton, N.J.: Educational Testing Service, August 1968.

Lannholm, Gerald V., and Schrader, William B. *Predicting graduate school success: An evaluation of the effectiveness of the Graduate Record Examinations.* Princeton, N.J.: Educational Testing Service, 1951. Out of print. (ETS Archives Microfiche #474)

Madsus, George F. *The development and use of expectancy tables for the Graduate Record Examinations Aptitude Test* (GRE Special Report 66-1). Princeton, N.J.: Educational Testing Service, April 1966.

Olsen, Marjorie. *The predictive effectiveness of the Aptitude Test and the Advanced Biology Test of the GRE in the Yale School of Forestry* (Statistical Report Number 55-6). Princeton, N.J.: Educational Testing Service, 1955. Out of print.

Reilly, Richard R. *Critical incidents of graduate student performance* (GRE Board Research Report 70-5). Princeton, N.J.: Educational Testing Service, June 1974a (ERIC Document Reproduction Service No. ED 058 313). Previously published as GRE Board Technical Memorandum No. 1, April 1971.

Reilly, Richard R. *Factors in graduate student performance* (GRE Board Professional Report 71-2P). Princeton, N.J.: Educational Testing Service, July 1974b (ERIC Document Reproduction Service No. ED 096 862). Previously published as Research Bulletin RB-74-2, February 1974.

Reilly, Richard R., and Jackson, Rex. *Effects of empirical option weighting on reliability and validity of the GRE* (GRE Board Professional Report 71-9P). Princeton, N.J.: Educational Testing Service, July 1974 (ERIC Document Reproduction Service No. ED 069 738). Previously published as Research Bulletin RB-72-38, August 1972.

Rock, Donald A. *The prediction of doctorate attainment in psychology, mathematics, and chemistry* (GRE Board Preliminary Report). Princeton, N.J.: Educational Testing Service, August 1972 (ERIC Document Reproduction Service No. ED 069 664). Later published as GRE Board Research Report 69-6aR, June 1974.

Rock, Donald A. *The "test chooser": A different approach to a prediction weighting scheme* (GRE Board Professional Report 70-2P). Princeton, N.J.: Educational Testing Service, November 1974.

Rock, Donald A. *The identification of population moderators and their effect on the prediction of doctorate attainment* (GRE Board Professional Report 69-6bP). Princeton, N.J.: Educational Testing Service, February 1975.

Sharon, Amiel T. *Test of English as a Foreign Language as a moderator of Graduate Record Examinations scores in the prediction of foreign students' grades in graduate school* (GRE Board Professional Report 70-1P). Princeton, N.J.: Educational Testing Service, June 1974 (ERIC Document Reproduction Service No. ED 058 304). Previously published as Research Bulletin RB-71-50, September 1971.

Bibliography—Part B

The validity studies in the bibliography below were analyzed in "Predicting Success in Graduate Education," an article written by the author of this paper and published in *Science*, 1974, Vol. 183, No. 4122. The studies concern the period 1952-1972. About half were published, and half were institutional reports or theses.

Alexakos, Constantine E. *The Graduate Record Examinations: Aptitude Tests as screening devices for students in the College of Human Resources and Education.* Unpublished report, West Virginia University, 1967. Reported by G.V. Lannholm in GRE Special Report 68-1. Educational Testing Service, Princeton, New Jersey, 1968.

Besco, Robert O. *The measurement and prediction of success in graduate school.* Ph.D. dissertation, Purdue University, 1960. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

Borg, Walter R. GRE aptitude scores as predictors of GPA for graduate students in education. *Educational and Psychological Measurement* 1963, 23 (2): 379-389.

Capps, Marian P., and Decosta, Frank A. Contributions of the Graduate Record Examinations and the National Teacher Examinations to the prediction of graduate school success. *Journal of Educational Research,* 1957, 50 (5): 383-389.

Clark, Henry. *Graduate Record Examination correlations with grade-point averages in the Department of Education at Northern Illinois University, 1962-1966.* Unpublished Master's thesis, Northern Illinois University, June 1968.

Conway, Sister Madona Therese. *The relationship of the Graduate Record Examination results to achievement in the Graduate School at the University of Detroit.* Unpublished Master's thesis, University of Detroit, 1955. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

Creager, John A. *A study of graduate fellowship applicants in terms of Ph.D. attainment.* Technical Report No. 18, Office of Scientific Personnel, National Academy of Sciences—National Research Council, Washington, D.C., 1961.

Creager, John A. *Predicting doctorate attainment with GRE and other variables.* Technical Report No. 25, Office of Scientific Personnel, National Academy of Sciences—National Research Council, Washington, D.C., 1965.

Dawes, Robyn M. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist,* 1971, 26 (2): 180-188.

Duff, Franklin L., and Aukes, Lewis E. *The relationship of the Graduate Record Examination to success in the Graduate College* (a supplementary comparative analysis of eight previously reported studies). Bureau of Institutional Research and Office of Instructional Research, University of Illinois, October 1966.

Eckhoff, Constance M. Predicting graduate success at Winona State College. *Educational and Psychological Measurement,* 1966, 26 (2): 483-485.

Ewen, Robert B. The GRE psychology test as an unobtrusive measure of motivation. *Journal of Applied Psychology,* 1969, 53 (5): 383-387.

Florida State University, Office of Academic Research and Planning. *The prediction of grade-point average in graduate school at the Florida State University,* Parts I & II. Florida State University, December 1971.

Florida State University, Office of Institutional Research and Service. *Relationship between Graduate Record Examinations Aptitude Test scores and academic achievement in the Graduate School at Florida State University.* Florida State University, 1958.

Hackman, J. Richard, Wiggins, Nancy, and Bass, Allan R. Prediction of long-term success in doctoral work in psychology. *Educational and Psychological Measurement,* 1970, 30: 365-374.

Hansen, W. Lee. *Prediction of graduate performance in economics.* Department of Economics, University of Wisconsin, April 1970 (Mimeographed).

Harvey, Philip R. *Predicting graduate school performance in education.* Unpublished ETS report, 1963. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

King, Donald C., and Besco, Robert O. The Graduate Record Examination as a selection device for graduate research fellows. *Educational and Psychological Measurement,* 1960, 20 (4): 853-858.

Lannholm, Gerald V., Marco, Gary L., and Schrader, William B. *Cooperative studies of predicting graduate school success.* GRE Special Report 68-3. Educational Testing Service, Princeton, New Jersey, August 1968.

Law, Alexander. The prediction of ratings of students in a doctoral training program. *Educational and Psychological Measurement,* 1960, 20 (4): 847-851.

Lorge, Irving. "Relationship between Graduate Record Examinations and Teachers College, Columbia University, Doctoral Verbal Examinations." Letter to G.V. Lannholm, dated September 21, 1960. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

Madaus, George F., and Walsh, John J. Departmental differentials in the predictive validity of the Graduate Record Examinations Aptitude Tests. *Educational and Psychological Measurement,* 1965, 25 (4): 1105-1110.

Mehrabian, Albert. Undergraduate ability factors in relationship to graduate performance. *Educational and Psychological Measurement,* 1969, 29 (2): 409-419.

Michael, William B., Jones, Robert A., Al-Amir, Hudhail, Pullias, Calvin M., Jackson, Michel, and Goo, Valerie. Correlates of a pass-fail decision for admission to candidacy in a doctoral program. *Educational and Psychological Measurement,* 1971, 31 (4): 965-967.

Michael, William B., Jones, Robert A., and Gibbons, Billie D. The prediction of success in graduate work in chemistry from scores on the Graduate Record Examinations. *Educational and Psychological Measurement,* 1960, 20 (4): 859-861.

Newman, Richard J. GRE scores as predictors of GPA for psychology graduate students. *Educational and Psychological Measurement,* 1968, 28 (2): 433-436.

Office of Educational Research ... *matriculating in the years 1952-* ..., 1963. Reported by G.V. Lannholm in ... cational Testing Service, Princeton, New ... 9.

Olsen, Marjorie. *The predictive effectiveness of the Aptitude Test and the Advanced Biology Test of the GRE in the Yale School of Forestry.* Statistical Report 55-6, Educational Testing Service, Princeton, New Jersey, 1955. Out of print.

Roberts, Pamela T. *An analysis of the relationship between Graduate Record Examination scores and success in the Graduate School of Wake Forest University.* Master's thesis, Wake Forest University, August 1970.

Robertson, Malcolm, and Nielsen, Winnifred. The Graduate Record Examination and selection of graduate students. *American Psychologist,* 1961, 16 (10): 648-650.

Robinson, Donald W. *A comparison of two batteries of tests as predictors of first year achievement in the graduate school of Bradley University.* Ph.D. dissertation, Bradley University, 1957. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

Rock, Donald A. *The prediction of doctorate attainment in psychology, mathematics, and chemistry* (GRE Board Preliminary Report). Princeton, N.J.: Educational Testing Service, August 1972 (ERIC Document Reproduction Service No. ED 069 664). Later published as GRE Board Research Report 69-6aR, June 1974.

Roscoe, John T., and Houston, Samuel R. The predictive validity of GRE scores for a doctoral program in education. *Educational and Psychological Measurement,* 1969, 29 (2): 507-509.

Sacramento State College, Test Office. *An analysis of traditional predictor variables and various criteria of success in the Master's degree program at Sacramento State College for an experimental group who received Master's degrees in the spring 1968, and a comparable control group who withdrew from their programs.* Test Office Report 69-3, Sacramento State College, October 1969.

Shaffer, Julie, and Rosenfeld, Howard. *MAT-GRE prediction study—Initial results.* Intradepartmental memorandum, Department of Psychology, University of Kansas, March 1969.

Sistrunk, Francis. "The GREs as predictors of graduate school success in psychology." Letter to G.V. Lannholm, dated October 3, 1961. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

Sleeper, Mildred L. Relationship of scores on the Graduate Record Examination to grade point averages of graduate students in occupational therapy. *Educational and Psychological Measurement,* 1961, 21 (4): 1039-1040.

Tully, G. Emerson. Screening applicants for graduate study with the Aptitude Test of the Graduate Record Examinations. *College and University,* 1962, 38: 51-60.

University of Virginia, Office of Institutional Analysis. *Correlations between admissions criteria and University of Virginia grade-point averages, Graduate School of Arts and Sciences,* Fall 1964. University of Virginia, circa 1966. (Mimeographed)

Wallace, Anita D. *The predictive value of the Graduate Record Examinations at Howard University.* Unpublished Master's thesis, Howard University, 1952. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

White, Elizabeth L. *The relationship of the Graduate Record Examinations results to achievement in the Graduate School at the University of Detroit.* Unpublished Master's thesis, University of Detroit, 1954. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

White, Gordon William. *A predictive validity study of the Graduate Record Examinations Aptitude Test at the University of Iowa.* Unpublished Master's thesis, University of Iowa, June 1967. Reported by G.V. Lannholm in GRE Special Report 68-1, Educational Testing Service, Princeton, New Jersey, 1968.

Williams, John D., Harlow, Steven D., and Grab, Del. A longitudinal study examining prediction of doctoral success: Grade-point average as criterion, or graduation vs. non-graduation as criterion. *Journal of Educational Research,* December 1970, 64 (4): 161-164.